Exhibit L

DNA Research 9, 47–57 (2002)

Characterization of Size-Fractionated cDNA Libraries Generated by the *in vitro* Recombination-Assisted Method

Osamu Ohara,*[,1,2] Takahiro Nagase,[1] Gaku Mitsui,[1,†] Hiroshi Kohga,[1] Reiko Kikuno,[1] Shuichi Hiraoka,[3] Yu Takahashi,[4] Satoshi Kitajima,[4] Yumiko Saga,[5] and Haruhiko Koseki[2,3]

Department of Human Gene Research, Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan,[1] RIKEN Research Center for Allergy and Immunology, 1-7-22 Suehiro, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan,[2] Department of Molecular Embryology, Graduate School of Medicine, Chiba University, 1-8-1 Inohana, Chuo-ku, Chiba 260-8670, Japan,[3] Cellular and Molecular Toxicology Division, National Institute of Health Sciences, 1-18-1 Kamiyohga, Setagaya-ku, Tokyo 158-8501, Japan,[4] and Division of Mammalian Development, National Institute of Genetics, Yata 1111, Mishima 411-8540, Japan[5]

(Received 3 April 2002)

Abstract

We here modified a previously reported method for the construction of cDNA libraries by employing an *in vitro* recombination reaction to make it more suitable for comprehensive cDNA analysis. For the evaluation of the modified method, sets of size-selected cDNA libraries of four different mouse tissues and human brain were constructed and characterized. Clustering analysis of the 3′ end sequence data of the mouse cDNA libraries indicated that each of the size-fractionated libraries was complex enough for comprehensive cDNA analysis and that the occurrence rates of unidentified cDNAs varied considerably depending on their size and on the tissue source. In addition, the end sequence data of human brain cDNAs thus generated showed that this method decreased the occurrence rates of chimeric clones by more than fivefold compared to conventional ligation-assisted methods when the cDNAs were larger than 5 kb. To further evaluate this method, we entirely sequenced 13 human unidentified cDNAs, named KIAA1990–KIAA2002, and characterized them in terms of the predicted protein sequences and their expression profiles. Taking all these results together, we here conclude that this new method for the construction of size-fractionated cDNA libraries makes it possible to analyze cDNAs efficiently and comprehensively.

Key words: cDNA sequencing; expression profile; cDNA library construction; *in vitro* recombination; size-fractionation; clustering

1. Introduction

Although the draft human genome sequence is now publicly available, cDNA cloning is still an indispensable step in the functional analysis of human genes. Thus, much attention has been given to the development of an efficient method for the construction of high-quality cDNA libraries.[1–5] In this regard, we recently reported the use of an *in vitro* recombination reaction based on the integrase-excisionase system of bacteriophage λ as an alternative to the ligation reaction mediated by DNA ligase for cDNA library construction.[6] This *in vitro* commercially available recombination cloning (RC) system, known as the Gateway cloning system (Invitrogen, USA),[7] requires a set of recombination sites in each of the donor and acceptor DNAs. The recombination sites are classified as attB (25 bp), attP (200 bp), attL (100 bp) and attR (125 bp). The attB site specifically recombines with the attP site to form the attL and attR sites, while the attL site always recombines with attR sites to give attB and attP sites. An enzyme mix which catalyzes the recombination reactions between attB and attP or between attL and attR is called the BP or LR clonase enzyme mix, respectively. Consistent with this nomenclature, the recombination reactions between attB and attP and between attL and attR are called the BP and LR reactions, respectively. DNA clones yielded by the BP and LR recombination reactions are conventionally termed "entry clones" and "expression clones," respectively. Thus, we hereafter term clones carrying attL sites and attB sites as entry clones and expression clones, re-

Communicated by Michio Oishi

* To whom correspondence should be addressed. Tel. +81-438-52-3913, Fax. +81-438-52-3914, E-mail: ohara@kazusa.or.jp

† Present affiliation: Pharmaceutical Research Department, Sato Pharmaceutical Co., Ltd., Tokyo, Japan

spectively, in this report. The RC system allows us to transfer cDNA inserts from one vector to another on demand with ease, high efficiency, and high fidelity, which has greatly facilitated functional analysis of genes.[7] If we use cDNA clones as reagents for functional analysis of genes, a set of cDNA clones compatible with the RC system must be an attractive resource. This is why we developed a new directional cDNA library construction method assisted by RC.[6] Besides this, we found that the RC-assisted cDNA library construction is preferable to conventional ligation-assisted cloning (LC) methods in terms of cloning efficiency, bias, suppression of occurrence of chimeric clones.[6]

In this study, we modified the library construction method previously reported so that cDNA clones are size-fractionated and can be efficiently used for DNA sequencing and protein production. To evaluate the performance of this modified RC-assisted cDNA library construction method, we prepared sets of size-fractionated cDNA libraries from four different mouse tissues and human brain using the modified method, and characterized the cDNA clones randomly isolated from the libraries by sequencing the cDNA ends. Clustering analysis of the end sequence data demonstrated that the libraries were complex enough for comprehensive cDNA analysis and that the frequency of occurrence of chimeric clones yielded by this method was considerably lower than that yielded by conventional LC-assisted methods. Furthermore, cDNA clones from human adult brain were subjected to entire sequencing. In particular, because we have been interested in human unidentified long cDNAs encoding large proteins, we analyzed the sequence data of 13 human unknown long cDNAs (> 3.6 kb) derived from genes named KIAA1990–KIAA2002, *in silico* and characterized them in terms of expression profiles as in our previous studies.[5,8,9] The results indicated that the modified RC-assisted cDNA library construction method coupled with size-fractionation successfully enabled us to analyze cDNA clones more comprehensively than other conventional methods and thus is well suited for comprehensive cDNA analysis. More importantly, the resultant cDNA clones served as versatile and powerful reagents for functional analysis of genes because they can be used for RC.

2. Materials and Methods

2.1. Materials

Mouse tissue RNAs were obtained from the following sources: Adult brain from BALB/c mice (Slc, Japan); embryonic intestinal tract [18.5 days post coitum (dpc); the day when the vaginal plug was detected was designated as 0.5 dpc] from ICR mice (Slc, Japan); adult thymus from C57BL/6 mice (Cr Slc); and embryonic tail from naturally mated ICR mice (CLEA, Japan). In particular, embryonic tails were prepared as follows: To enrich the transcripts which are possibly involved in the somite segmentation, we collected tail fragments containing parts of somitic and presomitic mesoderm corresponding to the S1, S0, S-1 and S-2 of 11.5 dpc embryo.[10] In total, 400 embryonic tail fragments were pooled and used for the preparation of total RNA. Total RNA was prepared using Isogen (Nippon gene, Japan) or TRIzole (Invitrogen), and purified to yield poly(A)$^+$ RNA with an mRNA isolation kit (Miltenyi Biotech, Germany). Human brain poly(A)$^+$ RNA was purchased from Clontech (USA).

In addition to the attP pSP73 donor plasmid,[6] a down-sized attP pSPORT-1 donor plasmid was used as an acceptor vehicle of cDNAs in the BP reaction in this study. This donor plasmid was prepared by deleting a 1-kb portion of laq I^q (nucleotide residues 745 to 1723 in pSPORT-1) and the 200-bp flanking region downstream from attP2 site in the attP cassette region of the previous attP pSPORT-1 donor plasmid.[6] The attP1 and attP2 sites in the down-sized attP pSPORT-1 donor plasmid are located close to SP6 and T7 promoters, respectively. A new attR donor plasmid used for the LR recombination reaction, termed "destination vector" by Invitrogen, was generated using a pBC SK$^+$ vector (STRATAGENE, USA) by inserting an attR/ccdB cassette between *Xho* I and *Sac* I sites in the multiple cloning site, as described in the previous report,[6] in which an attR1 site is located close to T7 promoter.

2.2. cDNA library construction

cDNA synthesis from poly(A)$^+$ RNA was carried out according to the supplier's instructions in the SuperScript^{TM} Plasmid System for cDNA Synthesis (Invitrogen). The first-strand synthesis was primed with an attB2-dT adapter primer (5′-FgcGCACCA<u>CTTTGTACAAGAAAGCTGG</u>GCGGCCGC$(T)_{18}$-3′, where F, g, and c indicate a fluorescein group and G and C residues with a phosphorothioate group, respectively). The underline shows the attB2 site sequence just upstream from a *Not* I site followed by dT_{18}. After synthesis of the second-strand DNA, the resultant cDNA was ligated with an attB1 adapter (the upper strand, 5′-CGACGCGTACAAGTTTGTACAAAAAAGCAGGCTCTTC-3′; the lower strand, 5′-GAAGAGCCTGCTTTTTTGTACAAACTTGTACGCG-3′) and then size-separated into two fractions (1 kb–3 kb and > 3 kb) by agarose gel electrophoresis. After recovering the size-fractionated cDNAs from the agarose gel with β-agarase, the cDNAs were then subjected to the BP reaction with the down-sized attP pSPORT-1 donor plasmid (for human brain cDNAs) or the attP pSP73 donor plasmid (for mouse cDNAs) as described previously.[6] The products of the BP reaction were purified by phenol extraction followed by ethanol precipitation and then in-

troduced into *Escherichia coli* cells by electroporation. We routinely used ElectroMAX DH10B cells (Invitrogen) for electroporation, because they consistently gave an extremely high transformation efficiency of 1.0–1.5 $\times$ 10^{10} colony formation units/μg of pUC19. cDNA plasmids in the form of an entry clone were prepared from more than 10^6 transformants grown on agar plates containing ampicillin (50 μg/ml) at 30°C after incubation in 2 $\times$ YT medium for 2–3 hr at 37°C. The resultant plasmids were run on agarose gel in super-coiled form and size-separated as agarose blocks along the lane. The size-fractionated cDNA plasmids were recovered from the agarose blocks using β-agarase, purified by phenol extraction followed by ethanol precipitation, and then used for the second-round LR reaction with a *Bam*HI-linearized attR pBC destination vector as instructed by the supplier of LR clonase (Invitrogen). The products were purified by phenol extraction followed by ethanol precipitation and then introduced into *E. coli* DH10B cells by electroporation, as described for the first-round recombination reaction. The size-selected cDNA plasmids in the form of an expression clone were prepared from more than 10^6 transformants for each size fraction in the same way as described above. The recovered plasmids in super-coiled form were again purified by agarose gel electrophoresis. The plasmids in each fraction were re-introduced to *E. coli* cells and again recovered from colonies ($> 10^6$ transformants) grown on agar plates as described. The extracted plasmids were electrophoresed and those with the expected size were again retrieved from the agarose gel. This gel-purification step was repeated 2–3 times until the recovered plasmids were almost free of contaminating small plasmids as judged from their electrophoretic patterns in super-coiled form.

2.3. *DNA sequencing*

Plasmid cDNAs were routinely prepared using MultiScreen 96-well filter plates (Millipore, USA) essentially as instructed by the supplier (http://www.millipore.com/analytical/publications.nsf/docs/TN004). End sequencing reactions of cDNA clones were performed with an M13 forward or a reverse primer using a BigDye terminator cycle sequencing kits (v1.0, Applied Biosystems, USA). The resultant products were analyzed using ABI 373A/377 DNA sequencers (Applied Biosystems) or a RISA 384-capillary DNA sequencer (Shimadzu, Japan). In particular, the end sequences of mouse cDNA clones were obtained exclusively by the RISA DNA sequencer. For deduction of the entire sequence, a cDNA insert was excised from the vector by digestion with *Xho* I and *Not* I, and purified by agarose gel electrophoresis. The recovered cDNA insert was self-ligated and then sheared by sonication. When a cDNA insert could not be excised from the vector with *Xho* I and *Not* I, the whole cDNA plasmid was subjected to sonication. The resulting fragments of 700–1000 bp were blunt-ended, retrieved by agarose gel electrophoresis and then ligated with dephosphorylated, *Sma* I-digested M13mp18 vector. We routinely analyzed 16 randomly isolated shotgun clones per 1 kb of cDNA insert. If gaps remained, they were filled by sequencing of the PCR products covering the gaps.

2.4. *Computer sequence analyses*

For the clustering analysis of the end sequences of mouse cDNA clones analyzed here, we first searched all the end sequences for repetitive elements to mask by RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html). Then, the resultant sequences were subjected to BLAST search[11] against a merged database which contained mouse mRNA sequences excluding expressed sequence tags (ESTs) extracted from GenBank (release 127.0) and the 3′-end sequences obtained in this study. When a query end sequence shared more than 90% sequence identity and had a score of more than 235 with any other sequences in the database by the BLAST search, the query and the hit sequences were grouped. The masked nucleotides were omitted from the sequence identity calculation. If a single sequence was found in common in any of the two different groups, the groups were merged into a single cluster. CAP2 was applied to assemble the sequences in each cluster and align them to make contigs.[12]

To identify the chimeric clones, 3′- and 5′-end sequence data of human brain cDNA clones, yielded by either the LC- or RC-assisted method, were first analyzed by clustering together with publicly available human cDNA sequences and the accumulated human cDNA end-sequence data at Kazusa DNA Research Institute (approximately 230,000 sequences) mainly according to the method described above. When 3′- and 5′-end sequences of a single clone were grouped into different clusters, we suspected the clone to be chimeric. The end sequences of the suspected clone were then mapped along the draft human genome sequence (ftp://ncbi.nlm.nih.gov/genomes/H_sapiens/) by BLAST. When an end sequence could be aligned to any genome contigs with higher than 99% sequence identity by BLAST at a threshold e-value of 10^{-50}, we assigned the chromosome number for the end sequence. If 3′- and 5′-end sequences were mapped on the genome contigs derived from the different chromosomes, we classified the clone as a chimeric clone.

To select cDNA clones to be entirely sequenced, the 3′- and 5′-end sequences of human brain cDNA clones were searched by BLAST against the GenBank database (release 123.0) excluding ESTs and genomic sequences. When the query end sequence did not have highly similar sequences in the database, that is, with the scores less than 500 and sequence identities lower than 85%, the end

sequence was categorized as an "unknown" sequence.

3. Results and Discussion

3.1. Construction of size-fractionated cDNA libraries by the two-round RC methods

In our method reported previously, cDNA clones were obtained in the form of an entry clone, where cDNAs were flanked by 100-bp attL sites.[6] This was simply due to the fact that the shortest recombination sites, i.e. attB sites, are most convenient to introduce into cDNA by synthetic oligonucleotides during cDNA synthesis. However, the presence of 100-bp attL sites flanking cDNAs was found to cause practical inconvenience in comprehensive cDNA analysis in some instances. In practice, end sequence data of cDNAs in the form of an entry clone must include the sequence of the attL site due to difficulty in designing an efficient sequencing primer within attL sites (http://www.invitrogen.com/Content/World/gatewayman.pdf), which results in decrease of the effective read length of the end sequences. In addition, the quality of the sequence data of entry clones are usually lower than that of conventional clones in our hands. On the other hand, the presence of an attL1 site upstream from the 5′-end of cDNA might interfere with protein production, particularly in the eukaryotic system, because it adds an artificial translation initiation site and/or termination site to cDNAs. Because we frequently adopt an *in vitro* transcription/translation assay for cDNA clones to analyze their protein products, this is a serious concern for us. Thus, we modified the previous method to yield cDNA clones in the form of an expression clone, where cDNAs are flanked by short recombination sites, i.e., attB sites of 25 bp. To yield cDNAs in the form of an expression clone in a single-round recombination reaction, attL sites must be added to cDNA ends. However, we found it practically impossible to prepare cDNA fragments flanked by attL sites without sacrificing simplicity and cloning efficiency. As the second choice, we thus decided to obtain the cDNA library in the form of expression clones through converting primary cDNA entry clones by the second-round LR reaction. A possible problem we anticipated in this process was a population bias of cDNA clones imposed by two rounds of the *in vitro* recombination reaction. In the previous study, we observed that the BP reaction caused a size bias in the cDNA population although the extent of the size bias was considerably smaller than that caused by LC.[6] We thus consider it critical to minimize the size-bias effect on the population of cDNAs in the second-round LR reaction. A simple solution to minimize the size-bias effect is to size-fractionate cDNAs prior to the second-round LR reaction. Because we have been extensively using size-fractionated cDNA libraries in our cDNA project,[5,8,9] the introduction of the size-selection step was not problematic at all. Using this modified cDNA library construction method, we generated sets of size-fractionated cDNA libraries from human adult brain and from mouse adult thymus, adult brain, embryonic intestine, and embryonic tail. These libraries were evaluated from a viewpoint of performance in comprehensive cDNA analysis as described below.

3.2. Characterization by agarose gel electrophoresis of size-fractionated cDNA libraries generated by RC-assisted method

To examine the quality of size-fractionated cDNA libraries yielded by the two-round recombination reactions, we first analyzed them as a mixture of cDNA plasmids by agarose gel electrophoresis (Figs. 1 and 2). Figure 1 shows the electrophoretic patterns of the size-fractionated human brain cDNA plasmids in the form of an entry clone (panel A), yielded just after the second-round recombination reaction (panel B), and finally obtained as an expression clone after gel purification (panel C). On the other hand, Fig. 2 displays the comparison of mixtures of the cDNA plasmids in the form of entry and expression clones after digestion with *Hin*dIII (panel A) or *Not* I (panel B) by agarose gel electrophoresis. In the previous report, we reported that large-size cDNA fractions contained a small but considerable amount of plasmid dimers.[6] Because some plasmids are known to have a tendency to become dimers spontaneously in *E. coli* cells and, in fact, plasmid dimers are sometimes found in conventional LC-assisted cDNA library, the occurrence of plasmid dimers in cDNA library was not so surprising. However, if amounts of contaminating plasmid dimers increase significantly, the performance of these size-fractionated libraries will be greatly reduced. Figure 1B shows that the second-round recombination reaction produced a considerable amount of plasmids with about a half of the size of the parental entry clones and the amount of these small plasmids increased with increase in size of cDNA clones included in the fractions. These small plasmids were likely segregated out of plasmid dimers. After removing these small plasmids in the second-round LR reaction products by purification on agarose gel, the size-fractionated cDNA libraries in the form of expression clones were ready to use for characterization (Fig. 1C).

In cDNA clones thus generated, the *Not* I and *Hin*dIII sites occur only once at the 3′-end of cDNA inserts originating from attB2-dT primer and at the 5′-flanking vector region in either vector, respectively. Because *Not* I is a rare cutter and thus rarely cut cDNA inserts internally, digestion of cDNA plasmids with *Not* I was expected to convert them to a linear form unless cDNA inserts had no internal *Not* I site. On the other hand, since *Hin*dIII sites are likely to occur more frequently in cDNAs than *Not* I sites, the electrophoretic patterns of plasmids after diges-

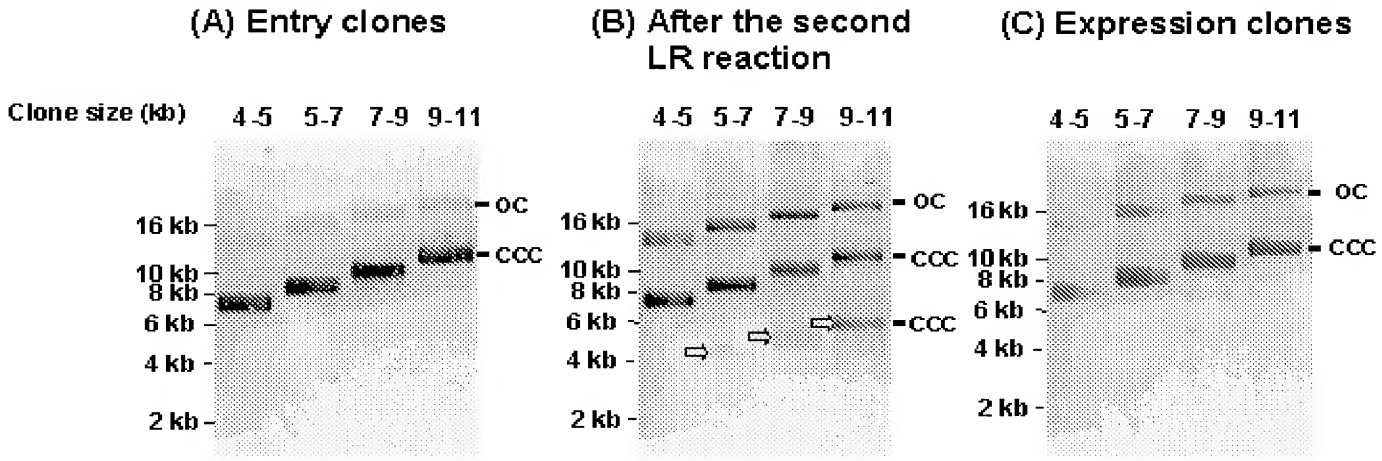

Figure 1. Electrophoretic patterns of size-fractionated human brain cDNA plasmids in super-coiled form. This figure shows 0.7% agarose gel images of a set of primary cDNA libraries in the form of entry clones (panel A), a set of the second-round recombination reaction products of the primary cDNA libraries recovered from transformants (more than 10^6 transformants for each size-fractionated cDNA library; panel B), and a set of size-fractionated cDNA libraries finally obtained after purification of cDNA plasmids shown in panel B on agarose gel (panel C). The cDNA plasmids shown in the panels B and C are expression clones. cDNA plasmids appeared as two bands depending on their physical forms [OC, open-circular form; CCC, covalently closed-circular (super-coiled) form]. The major forms of plasmids in respective bands are indicated at the right side of the gel images. In panel B, arrows indicate super-coiled cDNA plasmids segregated from plasmid dimers by the second-round recombination reaction. The positions of super-coiled DNA size markers are shown at the left side of each gel image. DNAs were visualized by staining with SYBR-Green I.

tion with *Hin*dIII are thought to give "fingerprints" of the size-fractionated cDNA libraries, which reflect the population of cDNA clones depending on their *Hin*dIII restriction maps. The restriction digestion with either *Not* I or *Hin*dIII was thus expected to convert plasmid dimers to two linear cDNA plasmids as long as the cDNAs do not contain an internal restriction site. The arrows in Fig. 2 indicate the bands of linearized cDNA plasmid monomers segregated from plasmid dimers by restriction digestion. The gel images indicated that the second-round recombination reaction considerably reduced the amount of cDNA plasmid dimers, evidently in the size-fractionated cDNA libraries of large size, probably because most of the cDNA plasmid dimers could not be propagated as plasmid dimers by the LR reaction (Fig. 2). Thus, the second-round recombination reaction certainly improved the quality of the size-fractionated libraries in this respect. Importantly, except for the decrease in the amount of cDNA plasmid dimers, the electrophoretic patterns of mixtures of the size-fractionated entry and expression cDNA plasmids after restriction digestion looked quite similar (Fig. 2). These results thus indicated that the population bias resulting from the second-round LR reaction was not drastic, if any, except for the loss of plasmid dimers.

In addition to plasmid dimers, we previously described that the cDNA libraries generated by the *in vitro* recombination reaction contained two additional types of clones with unexpected structures in small amounts (1% or less).[6] Although plasmids cloned by this method were expected to have only one *Not* I site at the 3′-end of cDNA insert, unless cDNA insert contains *Not* I site internally, the first type of clone could not be digested with *Not* I at all whereas the second type of clone released the entire part of cDNA insert from the vector upon restriction digestion with *Not* I. Sequence analysis of these peculiar clones revealed that: (a) the first-type cDNA originated from an attB2-dT primer which had a wrong *Not* I sequence accidentally; (b) the second-type cDNA was a 5′-end-to-5′-end chimera generated probably during attB1 adapter ligation.[6] We found that the second type of peculiar clones was generated by intermolecularly 5′-end-to-5′-end ligated cDNA fragments, where one of the cDNA fragments carried an attB1 adapter at the 3′-end. Because the 5′-end of the attB2-dT primer was not phosphorylated, this type of cDNAs is not expected to be generated using unphosphorylated attB1 adapter unless the 5′-end of the attB2-dT primer remains intact. In fact, we noticed that the attL1 site of the second-type peculiar cDNAs in the form of an entry clone was frequently followed by a truncated attB2 sequence. Although we used a modified attB2-dT primer with a bulky fluorescein group followed by phosphorothioate groups to attenuate degradation of the attB2-dT primer to some extent, Fig. 2 suggests that the second type of peculiar clones could not be excluded from the cDNA library because the bands of the vector size were seen after digestion of cDNA plasmids with *Not* I. These results indicate that the frequency of occurrence of the second type of peculiar plasmids was not affected much by the introduction

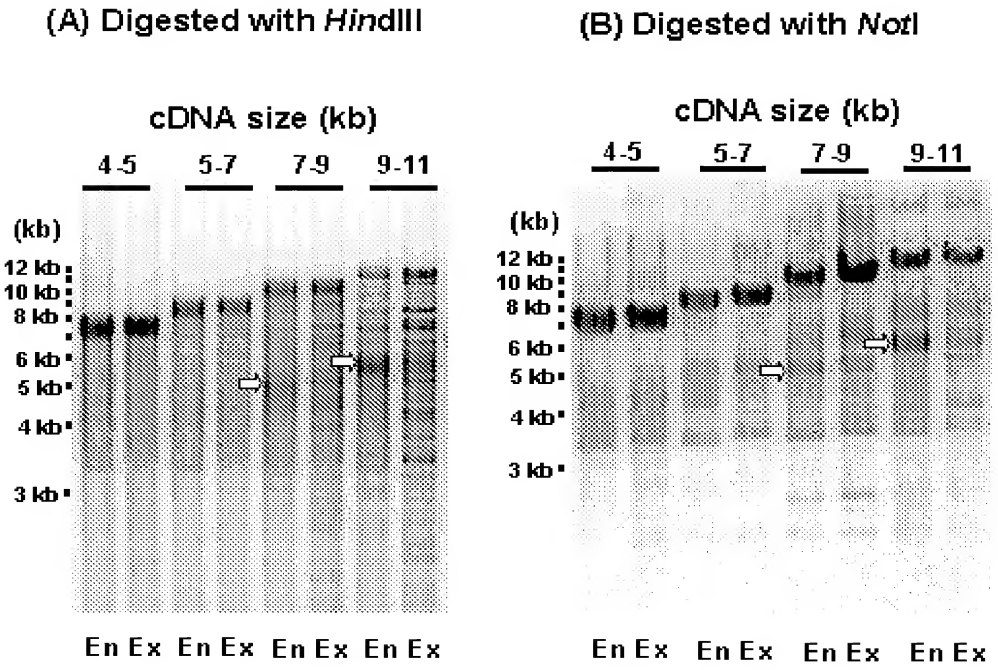

Figure 2. Comparison of electrophoretic patterns of size-fractionated human brain cDNA plasmids in the forms of entry and expression clones after restriction digestion. Panels A and B show the electrophoretic patterns of sets of the size-fractionated human brain cDNA plasmids in the forms of entry clones (En) and expression clones (Ex) after digestion with *Hin*dIII and *Not* I, respectively. Because the *Not* I site rarely occurs within cDNA inserts in general, the major band in each lane of panel B contained a linear cDNA plasmid digested at the *Not* I site in the attB2-dT primer. Minor bands seen in panel B contained internally *Not* I-digested cDNA plasmids, cDNA plasmids that could not be digested with *Not* I because of wrong sequences at *Not* I site in attB2-dT primer, the segregated plasmids from their parental plasmid dimers, and cDNA inserts and the vectors excised by digestion of artificial clones with *Not* I. Arrows indicate the positions of bands containing plasmids segregated from plasmid dimers. Except for changes in signal intensity of bands originating from plasmid dimers, the electrophoretic patterns in lanes En and Ex of each set of size-fractionated cDNA plasmids were not drastically different from each other. The vector portion of the entry clones, the down-sized pSPORT-1 in this case, was slightly smaller than the pBC vector used for the expression clones. DNAs were detected by fluorescence staining with SYBR-Green I. The positions of DNA size markers are shown at the left side of each gel image.

of the second-round recombination reaction. Because the first type of peculiar clones were generated by impurities in the preparation of the attB2-dT adapter primer, the occurrence rate of this type of peculiar clones was also unlikely to be affected by the second-round LR reaction.

3.3. Characterization of size-fractionated cDNA libraries by end sequencing of cDNA clones

The next concern to be addressed was the complexity of the resultant cDNA libraries. The results described above suggested that the introduction of the second-round recombination reaction did not drastically affect the population of cDNA clones as long as the cDNAs were size-fractionated prior to the second-round LR reaction. However, because the complexity of cDNA libraries is a critical concern in comprehensive cDNA analysis, the complexity of the size-fractionated cDNA libraries should be evaluated more directly by another method. We thus characterized mouse cDNA clones randomly isolated from size-fractionated libraries from different sources by sequencing at the 3′ end. The end sequences thus obtained were subjected to computer clustering analysis to determine how complex the resultant cDNA libraries were. The clustering analysis was performed using the 3′-end sequences of cDNA clones isolated from each size-fractionated cDNA library together with known mouse cDNA sequences in the public database. In this analysis, any cDNA clone with a 3′-end sequence which appeared only once in each set of 3′-end sequences was designated as a singleton. The clustering data were used to evaluate these libraries in terms of the occurrence rates of singletons and independent clones (i.e., singletons plus contigs) among the sequenced clones in each size-fractionated cDNA library. Since the occurrence frequency of uncharacterized cDNA clones in the public databases was also our concern in a practical sense, mouse cDNA sequences in the public database were included in this clustering analysis performed to check the novelty of the clones. For interpretation of the clustering data, we kept in mind that the occurrence rate of singletons and independent clones inevitably depends on the number of sequenced clones and

Table 1. Evaluation of complexities of size-fractionated mouse cDNA libraries generated by *in vitro* recombination method.

Brain

size (kb)	Occurrence rate of singletons (%)[1]	Occurrence rate of independent clones (%)[2]	Total number of 3′-end sequences analyzed
7-9	32.7 (30.9)	69.0 (44.5)	3953
5-7	32.0 (30.4)	72.7 (50.2)	3426
3-5	38.5 (36.5)	88.3 (68.4)	950
1-3	37.9 (27.8)	79.1 (57.1)	997

Thymus

size (kb)	Occurrence rate of singletons (%)[1]	Occurrence rate of independent clones (%)[2]	Total number of 3′-end sequences analyzed
6-8	17.7 (15.7)	60.5 (34.6)	1029
4-6	25.5 (22.8)	71.3 (46.5)	1743
3-4	27.7 (25.1)	77.5 (55.3)	1562
1-3	29.1 (25.7)	74.0 (60.5)	1812

Embryonic intestine

size (kb)	Occurrence rate of singletons (%)[1]	Occurrence rate of independent clones (%)[2]	Total number of 3′-end sequences analyzed
>11	11.2 (10.4)	41.2 (23.8)	936
9-11	18.1 (16.7)	55.4 (33.2)	1001
7-9	20.8 (18.3)	60.1 (36.5)	2538
5-7	25.5 (23.7)	71.1 (45.4)	2427

Embryonic tail

size (kb)	Occurrence rate of singletons (%)[1]	Occurrence rate of independent clones (%)[2]	Total number of 3′-end sequences analyzed
5-7	20.7 (19.0)	72.4 (47.0)	764
4-5	22.0 (20.1)	81.3 (50.8)	867
3-4	25.7 (23.6)	77.0 (56.1)	1567
2-3	27.4 (25.6)	66.1 (56.0)	2427
1-2	20.4 (14.4)	48.2 (37.8)	2826

1) The occurrence rates of singletons were calculated by dividing the number of 3′-end sequences appearing only once with the number of the total 3′-end sequences analyzed in each library. The number in parenthesis is the occurrence rate of singletons whose 3′-end sequences are not found in public databases as a cDNA entry by homology search of their terminal sequences.
2) The occurrence rates of independent clones were calculated by dividing the number of singletons plus the number of contigs by the total number of 3′-end sequences analyzed. The number in parenthesis is the occurrence rate of the independent clones whose 3′-end sequences are "unknown" as described in the text.

that the number of cDNA clones analyzed for each library varied from 764 to 3953, as shown in Table 1. Among the cDNA libraries examined, the set of size-fractionated cDNA libraries derived from brain exhibited the highest frequencies of occurrence of singletons as well as independent clones (Table 1). In most of the sets of the size-fractionated libraries, the occurrence rate of singletons and independent clones decreased with increasing cDNA size. This is probably due to the low complexity of the mRNA species in these size ranges and to the fact that *in vitro* reverse transcription cannot make a complete copy of extremely long mRNA with high efficiency. Interestingly, cDNA clones for mouse laminin $\alpha 5$ chain (GenBank accession no. U37501) occurred 103 times in a total of 936 clones of the mouse embryonic intestine cDNA library containing cDNAs larger than 11 kb. The presence of such a highly abundant cDNA species in a particular size range thus made statistical interpretation of the clustering data somewhat complicated. However, the statistical clustering data of the mouse cDNAs from various libraries supported the idea that the modified method assisted by two-round RC could successfully generate libraries with enough complexity for comprehensive cDNA analysis.

In the previous paper, we described that the RC method considerably reduced the frequency of occurrence of chimeric clones.[6] Because the human draft genome sequence data are publicly available at present, we could confirm this using the accumulated sequence data of human brain cDNA clones at both of their ends. By the clustering analysis, a cDNA clone with its 3′- and 5′-end sequences clustered into different clusters was suspected to be a candidate chimeric clone. If the end sequences of the suspected clone were mapped on different chromosomes, this clone was conclusively identified as a chimeric clone. Because the end sequences of all the sus-

Table 2. Comparison of frequency of occurrence of chimeric clones in size-fractionated cDNA libraries generated by *in vitro* recombination method and ligation method.

cDNA size	Occurrence rate of candidates of chimeric clones[1]	Occurrence rate of real chimeric clones in the suspected clones[2]	Frequency of occurrence of chimeric clones[3]
RC-based libraries			
7-9	4.7 % (26/559)	15.0 % (3/20)	0.7%
5-7	6.3 % (22/351)	0.0 % (0/15)	0.0%
4-5	8.0 % (27/339)	5.9 % (1/17)	0.5%
2-3	6.0 % (19/319)	6.3 % (1/16)	0.4%
1-2	2.3 % (98/4257)	28.3 % (15/53)	0.6%
LC-based libraries			
7-9	13.5 % (85/630)	56.3 % (18/32)	7.6%
5-7	9.7 % (315/3238)	44.6 % (74/168)	4.3%

1) The frequencies were obtained by dividing the number of suspected clones by the number of clustered clones which had sequence data of both the 5′ and 3′ ends. The actual numbers used for the calculation are given in parentheses.
2) The frequencies were obtained by dividing the number of the suspected clones whose end sequences were mapped on different chromosomes by the number of the suspected clones which had genome mapping information for both the 5′ and 3′ ends. The actual numbers used for the calculation are given in parentheses.
3) The frequencies were obtained by multiplying the occurrence rate of the suspected clones from clustering analysis with that of real chimeric clones in the suspected clones.

pected clones could not be mapped on the draft human genome sequence currently available, the number of the clones which could be mapped at both 5′ and 3′ ends is also shown in Table 2. We assumed that the occurrence rate of real chimeric clones is obtained by multiplying the occurrence rate of the suspected clones in the sequenced clones with that of the chimeric clones finally identified in the suspected clones. As shown in Table 2, the occurrence rate of chimeric clones in each size-fractionated cDNA library generated by the RC method is always below 1%. In contrast, chimeric clones appeared at a rate of 4.3% or 7.6% in the LC-assisted libraries with the cDNA size between 5–7 kb and 7–9 kb, respectively. These results confirmed our previous observation and clearly demonstrated the advantage of the RC method over the conventional LC method.

3.4. Characterization of long cDNA clones isolated from human brain by entire sequencing

To finally evaluate the performance of the size-fractionated cDNA libraries, we actually determined entire sequences of human brain cDNAs isolated from the RC-assisted libraries. Because we have sequenced human brain long cDNAs by the conventional LC method for 7 years,[5,8,9] it was important to know whether or not the entire sequencing of these cDNA clones obtained by the RC method could be done as efficiently as that of the cDNA clones generated by LC. In conclusion, we confirmed that shotgun sequencing of the RC-generated clones could be carried out without difficulty; even when cDNA inserts could not be readily excised with restriction enzymes, shotgun sequencing of the whole plasmid worked well only if the number of shotgun clones analyzed was slightly increased. In practice, 13 cDNA clones (KIAA1990–KIAA2002) were selected for sequencing in their entirety because they could produce proteins with an apparent molecular mass larger than 50 kDa in an *in vitro* transcription/translation system and have "unknown" end sequences as described above.[5] The screening by the *in vitro* transcription/translation system revealed that cDNA clones yielded by RC in the form of an expression clone could produce proteins *in vitro* as efficiently as those obtained by LC.

The sequence features of the 13 clones are listed in Fig. 3 and Table 3. Regarding three cDNA clones (KIAA1999–KIAA2001), multiple protein-coding sequences (CDSs) were detected in a single cDNA by GeneMark analysis.[14] Thus, we carefully checked whether the observed interruption of CDSs was spurious or not. The coding split of the cDNA clone for KIAA1999 was found to be spurious by experimental examination of the direct sequencing of the major reverse transcription-coupled polymerase chain reaction (RT-PCR) products and thus this cDNA sequence was revised according to the RT-PCR results. For this cDNA clone, the revised sequence, not the actual cloned cDNA sequence, was deposited to GenBank/EMBL/DDBJ databases and used for the prediction of CDSs. Although the cDNA clone for KIAA2001 triggered a CDS split alert by GeneMark analysis,[14] this alert turned out to be false, at least for mRNAs in brain, since the sequence obtained from the major product amplified by RT-PCR was the same as that of the cDNA clone. For information purposes, the difference between the cloned cDNA and the revised cDNA sequences is shown on our web site, HUGE (http://www.kazusa.or.jp/huge).[15] In contrast to the cases described above, for the cDNA sequence for KIAA2000, we could not obtain RT-PCR products for the regions spanning the predicted CDS interruption in sufficient amount and purity for sequencing. Thus, only

Table 3. Information of newly identified genes.

Gene number (KIAA)	Accession number[b]	cDNA length (bp)[c]	ORF length (amino acid residues)	Chromosomal location[d]	Results of homology search against nr database[a]				
					nr ID	aa. res.	% Identity	% Coverage[f]	Definition
1990	AB082521	7999	823	16	AF026934	878	92	99	pyruvate dehydrogenase phosphatase regulatory subunit precursor, mRNA, complete cds - *Bos taurus*
1991	AB082522	7811	712	11	none				
1992	AB082523	6583	853	2	AB070036	724	54	86	testis cDNA clone:QnA-12007, full insert sequence - *Macaca fascicularis*
1993	AB082524	6545	532	9	AF395817	527	35	38	NAC1 protein mRNA, complete cds - human
1994	AB082525	4815	1117	13	AF201285	1057	85	94	TSC22-related inducible leucine zipper 1b (TILZ1b) mRNA, complete cds - mouse
1995	AB082526	5492	1011	14	P51954	774	42	26	SERINE/THREONINE-PROTEIN KINASE NEK1 (EC 2.7.1.-) - mouse
1996	AB082527	3682	437	10	P07106	533	81	99	ENDOZEPINE-RELATED PROTEIN PRECURSOR - *Bos taurus*
1997	AB082528	3955	1194	11	AB041881	4306	94	100	dnchc2 mRNA for cytoplasmic dynein heavy chain, complete cds - rat
1998	AB082529	8407	607	5	P97433	1693	77	97	RHO-INTERACTING PROTEIN 2 - mouse
1999[e]	AB082530	8213	1275	5	Q09743	1309	31	44	PROTEIN STI1 - *Schizosaccharomyces pombe*
2000	AB082531	7784	699	3	S33816	451	34	48	kinesin light chain isoform 4 - sea urchin
2001	AB082532	4356	689	X	AB049834	325	31	33	PEG10 mRNA for paternally expressed gene 10 ORF1, complete cds - human
2002	AB082533	8282	764	15	AK024793	329	100	43	FLJ21140 fis, clone CAS07548 - human

a) Homology search was performed by FASTA against the nr database. The homologous protein with the highest score was listed, when it satisfied the following conditions: i) the aligned region exceeded 200 amino acid residues, and ii) percent identity in the aligned region was 30% or greater.
b) Accession numbers of DDBJ, EMBL and GenBank databases.
c) Values excluding poly(A) sequences.
d) Chromosome numbers were determined from the results of BLAST search of cDNA clones against the human draft genome sequence (ftp://ncbi.nlm.nih.gov/genomes/H_sapiens/) unless specified. The chromosomal location highlighted by an asterisk was fetched from the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene/).
e) cDNA and ORF lengths were revised by direct analysis of the RT-PCR products.
f) The values are the ratio of the length of the aligned region to the original length of the query sequence, expressed as a percentage.

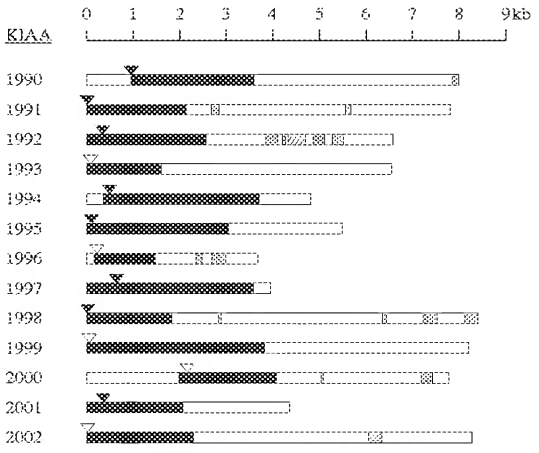

Figure 3. Physical maps of cDNA clones analyzed. The physical maps shown here were constructed from the sequence data of respective cDNA clones or, when necessary, from the combination of cDNA clones and RT-PCR products. The horizontal scale represents the cDNA length in kb, and the gene numbers corresponding to respective cDNAs are given on the left. The ORFs and untranslated regions are shown by solid and open boxes, respectively. Regarding KIAA2000, only the largest CDSs predicted by GeneMark analysis are shown, even though multiple CDSs are predicted. Information on the multiple CDSs is available through our web site HUGE.[15] The positions of the first ATG codons with or without the contexts of the Kozak's rule are indicated by solid and open triangles, respectively.[16] RepeatMasker, which is a program that screens DNA sequences for interspersed repeats known to exist in mammalian genomes, was applied to detect repeat sequences in respective cDNA sequences (Smit, A. F. A. and Green, P., RepeatMasker at http://ftp.genome.washington.edu/RM/ RepeatMasker.html). Short interspersed nucleotide elements (SINEs) including Alu and MIRs sequences and other repetitive sequences thus detected are displayed by dotted and hatched boxes, respectively.

the longest CDS is shown in Fig. 3 even though it has multiple predicted CDSs by GeneMark analysis as described above. Figure 3 shows the open reading frames (ORFs) and the first ATG codons in respective ORFs of these 13 cDNAs using solid boxes and triangles, respectively. Repeat sequences analyzed by the RepeatMasker program are also displayed in Fig. 3. In conclusion, the average size of the 13 cDNA sequences reached 6.5 kb and that of the predicted CDSs corresponded to approximately 829 amino acid residues. Table 3 lists the lengths of inserts and the ORF lengths of the respective clones in addition to the results of the homology search against a non-redundant amino acid sequence database, nr (ftp://ncbi.nlm.nih.gov/blast/db/nr.z). As additional information on these KIAA genes, the chromosomal loci of genes were determined by comparing the cDNA sequences with human genome draft sequence (ftp://ncbi.nlm.nih.gov/genomes/H_sapiens/) (Table 3).

As additional information regarding the characteristics of these newly identified KIAA genes, the expression profiles in 10 human tissues (8 brain regions, spinal cord, fetal liver and brain) were determined by quantitative RT-PCR coupled with an ELISA as described previously.[6] The tissue expression profiles of them are shown in Fig. 4, which might provide us with some insights as to their physiological role in the future.

All the results described above are accessible through the HUGE protein database at http://www.kazusa.or.jp/huge.[15]

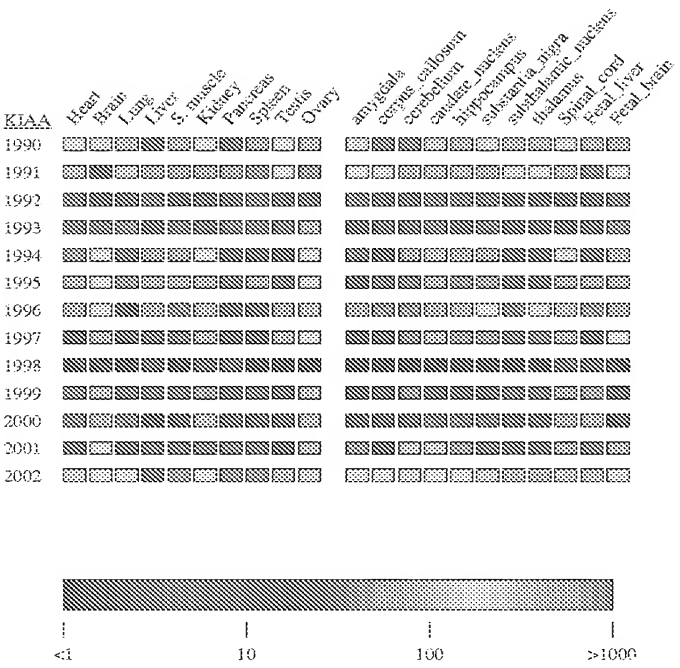

Figure 4. Expression profiles of 13 newly identified genes examined by RT-PCR ELISA. The tissue expression levels of the 13 human genes were analyzed using the RT-PCR ELISA according to the methods previously described in detail.[17] Gene names are given as KIAA numbers at the left side of each set of color codes. Tissue and brain region names are indicated above the top sets of color codes. A color conversion panel shown at the bottom is used for displaying mRNA levels as color codes. The mRNA levels are expressed in equivalent amounts (fg) of the authentic cDNA plasmids in 1 ng of starting poly(A)$^+$ RNAs. In addition, 10 tissues including 9 regions of the adult central nervous system (amygdala, corpus callosum, cerebellum, caudate nucleus, hippocampus, substantia nigra, subthalamic nucleus, thalamus, and spinal cord) and fetal brain were included in the expression profiling. As a control, mRNA levels in fetal liver were also examined.

4. Concluding Remarks

The results obtained in this study clearly demonstrated that the modified RC-assisted cDNA library construction is an efficient way to analyze cDNAs comprehensively. More importantly, cDNA clones yielded by this method are quite useful for functional analysis of genes because they can be easily converted to other plasmid forms (e.g., various types of expression plasmids) by RC in a single tube as instructed by Invitrogen. It is considered to be possible to further improve this method by combining this method with other existing technologies such as full-length enrichment methods and normalization/subtraction methods, depending on the purpose of each project.

Acknowledgements: This project was supported by a grant from the Kazusa DNA Research Institute and grants from Special Coordination Funds and for Scientific Research on Priority Areas (C) of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government. We thank Tomomi Tajino, Keishi Ozawa, Tomomi Kato, Kazuhiro Sato, Akiko Ukigai, Kazuko Yamada, Kiyoe Sumi, Takashi Watanabe, Sachiko Minorikawa, Kozue Kaneko, Naoko Shibano, Mina Waki, and Nobue Kashima for their technical assistance.

References

1. Maruyama, K. and Sugano, S. 1994, Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides, *Gene*, **138**, 171–174.
2. Edery, I., Chu, L. L., Sonenberg, N., and Pelletier, J. 1995, An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture), *Mol. Cell Biol.*, **15**, 3363–3371.
3. Carninci, P., Kvam, C., Kitamura, A. et al. 1996, High-efficiency full-length cDNA cloning by biotinylated CAP trapper, *Genomics*, **37**, 327–336.
4. Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R., and Siebert, P. D. 2001, Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction, *BioTechniques*, **30**, 892–897.
5. Ohara, O., Nagase, T., Ishikawa, K.-I. et al. 1997, Construction and characterization of human brain cDNA libraries suitable for analysis of cDNA clones encoding relatively large proteins, *DNA Res.*, **4**, 53–59.
6. Ohara, O. and Temple, G., 2001, Directional cDNA library construction assisted by the *in vitro* recombination reaction, *Nucleic Acids Res.*, **29**, e22.
7. Walhout, A. J., Temple, G. F., Brasch, M. A. et al. 2000, GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes, *Methods Enzymol.*, **328**, 575–592.
8. Nomura, N., Miyajima, N., Sazuka, T. et al. 1994, Prediction of the coding sequences of unidentified human genes. I. The coding sequences of 40 new genes (KIAA0001–KIAA0040) deduced by analysis of randomly sampled cDNA clones from human immature myeloid cell line KG-1, *DNA Res.*, **1**, 27–35.
9. Nagase, T., Kikuno, R., and Ohara, O. 2001, Prediction of the coding sequences of unidentified human genes. XXII. The complete sequences of 50 new cDNA clones which code for large proteins, *DNA Res.*, **8**, 319–327.
10. Pourquie, O. and Tam, P. P. 2001, A nomenclature for prospective somites and phases cyclic gene expression in the presomitic mesoderm, *Dev. Cell*, **1**, 619–620.
11. Altschul, S. F., Madden, T. L., Schaffer, A. A. et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–3402.
12. Huang, X. 1996, An improved sequence assembly program, *Genomics*, **33**, 21–31.
13. Ishikawa, K.-I., Nagase, T., Suyama, M. et al. 1998, Prediction of the coding sequences of unidentified human genes. X. The complete sequences of 100 new cDNA clones from brain which can code for large proteins *in vitro*, *DNA Res.*, **5**, 169–176.

14. Borodovsky, M., McIninch, J. D., Koonin, E. V., Rudd, K. E., Medigue, C., and Danchin, A. 1995, Detection of new genes in a bacterial genome using Markov Models for three gene classes, *Nucleic Acids Res.*, **23**, 3554–3562.
15. Kikuno, R., Nagase, T., Waki, M., and Ohara, O. 2002, HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project, *Nucleic Acids Res.*, **30**, 166–168.
16. Kozak, M. 1996, Interpreting cDNA sequences: some insights from studies on translation, *Mammalian Genome*, **7**, 563–574.
17. Nagase, T., Ishikawa, K.-I., Suyama, M. et al. 1998, Prediction of the coding sequences of unidentified human genes. XI. The complete sequences of 100 new cDNA clones from brain which code for large proteins *in vitro*, *DNA Res.*, **5**, 277–276.